DOCUMENT RESUME

ED 453 279 TM 032 816

AUTHOR	Taherbhai, Husein; Young, Michael James
TITLE	An Analysis of Rater Impact on Composite Scores Using the Multifaceted Rasch Model.
PUB DATE	2000-04-00
NOTE	22p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 25-27, 2000).
PUB TYPE	Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Evaluators; *High School Students; High Schools; *Interrater Reliability; Item Response Theory; Reading Tests; *Scores
IDENTIFIERS	*Rasch Model; *Weighting (Statistical)

ABSTRACT

This empirical study used data from the Reading: Basic Understanding section of the New Standards English Language Arts Examination. Data were collected for 3,200 high school students randomly selected from those who took the examination. The resulting sample had 16 raters who scored 200 students each, with each student rated by only 1 rater. The data were manipulated to form different weighted composite scores, which were then analyzed for rater effects, using the multifaceted Rasch model. Results indicate that main and interactive effects of raters and weighted composite scores can have varied effects on student ability estimates. Care in using weighted scores is suggested, and the use of simulated data is recommended to replicate empirical results with both the one-parameter and the two-parameter item response theory models. (SLD)

ED 453 279

An Analysis of Rater Impact on Composite Scores Using the Multifaceted Rasch Model

Husein Taherbhai
Learning Research and Development Center
University of Pittsburgh

Michael James Young
Harcourt Educational Measurement

TM032816

Paper presented at the annual meeting of the National Council of Measurement in Education, April 25-27, 2000, New Orleans, LA.

Summary

Constructed responses or open-ended tasks have seen a great insurgence in recent years. Since these tasks cannot be machine-scored, trained raters are used to score them. However, variability among raters cannot be completely eliminated and, therefore, rater effects cast doubts on the reliability of the study when they are not modeled. Besides rater effects, differentially weighted tasks/items that formulate composite scores can also have an effect in the estimation of student ability. These composite scores can have a compounding effect on student abilities when they interact with rater effects.

This empirical study uses data from the Reading: Basic Understanding section of the New Standards English Language Arts Examination. The data are manipulated to form different weighted composite scores, which are then analyzed for rater effects, using the multifaceted Rasch model.

Results indicate that main and interactive effects of raters and weighted composite scores can have varied effects on student ability estimates. Care in using weighted scores is suggested and simulated data are recommended to replicate empirical results both with the one-parameter and the two-parameters IRT models.

Inequitable situations in assigning grades can arise when examinees' composite scores are affected by differences in rater severity. Typically, testing programs subject raters scoring open-ended items to extensive training and quality control checks. However, in spite of the precautions used by testing programs for fair and equitable scoring, there may be instances when the behavior of raters "must be modeled and statistically controlled" (Linacre, 1993, p. 6) to provide for greater equity in the reporting of student scores. Since a rater can differ from other raters in terms of rater severity or leniency, differences can arise in assigning composite scores and the corresponding classifications of achieving a standard.

This situation is especially compounded in the case of scoring open-ended items where rater severity is confounded by item difficulty. In these circumstances, the impact of rater effects may be intensified when tasks are differentially weighted to form a composite score. Examinees may then receive grades that are not equitable due to the effects of rater severity and the differential weighting of the examinees' responses. For example, if open-ended (OE) questions are given more weight than the multiple-choice (MC) items, a student who does poorly on the open-ended section will suffer worse consequences in comparison to a peer who does well on the open-ended but does poorly on the multiple-choice items.

However, student performance is not only a function of differential weights assigned to his/her responses on the respective sections (MC or OE) but also on the contribution of each section to the total composite score. In the example above, the student who performs badly on a difficult OE section may come out ahead of his peer who does well on the OE section if, say, the OE section contributes only 10% to the student composite score while the remaining 90% is contributed by the MC section.

This study aims to analyze rater effects on complex structure of student composite scores (similar to AP structure) derived from the Reading: Basic Understanding section of the New Standards (NS) English Language Arts Examination. The study provides comparative analysis on student performance with and without the consideration of rater effects for differently weighted composite scores using the multifaceted Rasch model. The study will also examine the effects of raters on composite scores for student classifications based on cutpoints.

Design and Methodology

Various log linear models can be used to analyze the hypothesis of no rater effect in scoring the open-ended sections of the NS examinations. In Item Response Theory, the multifaceted Rasch model for ordered response categories (Linacre, 1989) can provide information on examinees, items, raters, and their interactions. The resulting probabilistic equation for a modified partial credit model (Masters & Wright, 1981) incorporating the different measurement facets (i.e., students, raters, and items) can be presented in logarithmic form as:

$$\log\left[\frac{P_{nijk}}{P_{nijk-1}}\right] = \beta_n - \delta_i - \lambda_j - \tau_{ijk} \tag{1}$$

where

P_{nijk} = probability of examinee n being rated k on item i, by rater j,

P_{nijk-1} = probability of examinee n being rated k-1 on item i, by rater j,

β_n = ability of examinee n,

δ_i = difficulty of item i,

λ_j = severity of rater j,

τ_{ijk} = difficulty of rater j in rating step k relative to step k-1 for item i.

The parameters of this model can be estimated using the FACETS program by Linacre (1989). A chi-square test of no difference among raters and an examination of the reliability of separation index will provide information as to whether the raters differed significantly across examinees or items. The reliability of separation index (*R*) obtained by the FACETS program is analogous to the traditional reliability indices such as KR-20 and coefficient alpha, in the sense that it reflects the ratio of true score variance to observed score variance (Engelhard, 1994).

The INFIT and OUTFIT indices, which together provide the standardized residuals, were examined for the identification of errant raters. Both statistics have expected values of 1.0 when the model fits the data. Each rater's fit statistics were examined, with acceptable fit ranging from 0.6 to 1.5 (Lunz, Wright, and Linacre's, 1990). Engelhard (1994) has found that these values provide a useful "rule of thumb" for substantive interpretations of overall rater behavior.

Examinees, Raters and the Instrument

Three thousand two hundred high school students were randomly selected from a total of 10,248 students who took Form C of the New Standards (NS) English Language Arts Examination (ELA). The resulting sample had 16 raters who scored a total of 200 students each. Each student, however, was rated by only one rater. Data collected pertained to the Basic Understanding section of the ELA Examination. This part of the examination consisted of 14 multiple-choice questions (MC) scored dichotomously (0,1) and one open-ended task (OE) which was scored on a 0 to 5 rubric.

Procedure

In order to analyze the impact of weights assigned to the questions with and without rater effects, six different composite scores were created. To facilitate comparison of the parameters under the different composites and apply the same cutpoint across the different composites, steps were taken to create composite scores that would not fluctuate beyond an upper limit value after being weighted. This was done by assigning different weights to the multiple-choice items and the open-ended task so as to produce a composite score that would not exceed "19." This baseline maximum score was chosen based on the unweighted composite score which would

range from 0 (no MC items correct and a minimum score of 0 on the OE item) to 19 (all 14 MC items correct and a maximum score of 5 on the OE item).

The first composite (75/25) consisted of no rater effects with weights of 1 assigned to the 14 MC items and the OE task. This provided a contribution of approximately 75% by the MC items and 25% by the OE task to the students' total composite score. The second composite (50/50) with no rater effects was 50% contribution to the composite scores by the MC items and 50% by the OE task. This implied a weight of 0.679 for the MC items and 1.900 for the OE task to obtain a maximum score of 19. Similarly, a composite (25/75) based on 25% contribution by the MC items and 75% contribution by the OE task to the composite scores was undertaken with weights = 0.339 and 2.850 for the MC items and the OE task, respectively. Composite scores 4, 5, and 6 were created with the same sets of 75/25, 50/50, and 25/75 weights, but included rater effects. These six different composite scores with the weights used to derive them are presented in Table 1.

Insert Table 1, here.

In order to establish a common metric for comparing the different composite scores, all of the MC items were first calibrated separately using the FACETS program to produce a Rasch item difficulty estimate for each item. The FACETS calibrations for each of the six composites were anchored to these MC parameter estimates when producing Rasch parameter estimates pertaining to student ability, rater severity, and the OE task difficulty. The Rasch ability estimates were found for each score point on each the six weighted composite scales. Cutpoints at the quartiles for the Rasch ability estimates of the baseline composite (i.e., no rater effects and weights of 1 for both the MC items and the OE task) were used to examine changes in student classifications across the different composites.

Results

Tables 2, 3, and 4 provide descriptive statistics for the multiple-choice total, the open-ended item, and Composites 1, 2, and 3 for the overall sample and for each rater.

The mean of the MC items was 10.49 with SD = 2.60. The OE task was graded on a 1-5 rubric and had a mean of 2.74 with SD = 0.87.The 14 MC items in the New Standards Basic Understanding (BU) cluster of the ELA examination correlated .39 with the open-ended task (OE). The 16 raters included in this study had scoring means with a low of 2.24 and SD = 0.98 (rater # 1181) to a high of 3.21 and SD = 0.96 (rater # 817) for the OE task.

<u>Insert Tables 2, 3, and 4, here.</u>

Table 5 shows the rater measures for Composites 4, 5, and 6. As can be seen, the most severe rater is #1407, while the least severe rater is #402. The chi-square test for no difference among raters was significant at the .01 level for the 75/25 condition, indicating substantial differences among raters in their rating behavior (χ^2 = 497.4, df=15). As would be expected the chi-squares were also significant for the other two composites (χ^2 = 1303.6, df=15; and χ^2 = 2353.8, df=15, for the 50/50 and the 25/750 composites, respectively). The reliability of the separation index also are very high (.97, .99, and .99 for the three composites, respectively) further indicating that the rater performances are indeed very different from one another.

<u>Insert Table 5, here.</u>

The rater fit indices for Composite 4 (75/25) are exemplary, with a low of 0.9 and a high of 1.2. Since FACETS incorporates the weights as a recurring response scored identically by the same rater, the raters became substantially more muted when the weighing of the OE task was increased to 1.90 and 2.85 in Composites 5 and 6. This would be expected because when the OE

item is weighted more than 1, raters seem to assign the same observed OE task scores across composites, and thus seem to score "holistically" (see Engelhard, 1994, for substantive meaning of fit indices for rater performance).

Insert Table 6, here.

The recovery of parameters also worsened with increasing weights assigned to the OE task. Once again, this was to be expected because rater individual fit indices, as measured by their INFIT and OUTFIT indices, become substantially muted as OE weights are increased. This has a compounding effect on the fact that MC items were anchored across composites, thus forcing parameter values that may not be the best estimates in conjunction with the other elements in the model.

Figure 1 depicts the effects of different weights on composite scores when no rater effects are involved. Student ability scores are very much higher under the 25/75 condition, i.e., when the OE task is weighed the most. This is indicative of the difficulty of the OE task relative to the MC items which forces an increase in the ability ratings of students because of the effect of the harder OE task. If the OE task had been easier than the MC items, a higher increase in the weights of the OE task with a corresponding fall in the weights of the MC items, would have created an opposite effect and the curves under conditions 25/75 and 75/25 would have changed places.

Insert Figure 1, here.

Figures 2 to 4 plot the composite scores against student ability scores (theta) for each of the three composites that include rater effects. As can be seen from the figures, rater discrepancy increases with increasing weights assigned to the OE task. This is understandable since

increasing OE weights compound the differences that already exist before weights are increased. However, the increase across the different conditions is not uniform, which probably is an artifact of the sample size. In each of the three rater-effect conditions, rater discrepancy increases with increasing composite raw score, indicating that raters are not in agreement with the student ability scores at higher composite scores, i.e., at the level where students tend to score higher on the OE task.

Insert Figures 2, 3, and 4, here.

Table 7 shows the consistency of classification of students with respect to cutpoints at the quartiles, for the adjusted and unadjusted ability estimates. The specific comparisons are shown for each set of weights (i.e., Composite 1 vs. Composite 4, Composite 2 vs. Composite 5, Composite 3 vs. Composite 6). Each of these sets of comparisons show a high level of consistency in classifying students with and without rater effects for the different composites. However, there were noticeable differences in the way students changed classification across the sets of weights: In the Composite 1 vs. Composite 4 comparisons all of the changes in student classification were downward. Overall, raters seem to be lenient (there is more downward movement than upwards in student classification) with one exception. For Composite 3 vs. Composite 6, there are more students who changed their above the Quartile 1 classification, indicating that on an average, raters are more severe at this cutpoint.

Insert Table 7, here.

Conclusion and Discussions

The impact of rater effects is well documented in literature (Engelhard, 1994, 1996; Linacre, 1989; Lunz, et al., 1990). The effect of weighing items differently compounds rater effects and further undermines the equity issue of the examination. It is evident that disparity among student scores get magnified with greater than 1 weights attached to open-ended tasks, especially when such discrepancies already exist because of rater effects. The consequence of weighing an open-ended task is also dependent on the how hard or easy the task is in relation to the other items in the test. For example, weighing an easy item has no other *statistical* purpose other than inflating most students' scores. On the other hand, further discrimination among the low and higher achieving student would ensue when hard items are weighted higher than their base weight of 1.

There are a number of testing programs that use proportional weighing corresponding to the predetermined contribution of the items to the students' composite scores[1]. Because weighted composite raw scores may not be the same for identical unweighted raw scores, two students with the same unweighted raw score would have different ability estimates.

For example, Table 8 considers the composite scores assigned to several students by rater #1117. The ability estimate of student # 13889, who had a score of 11 for Composite 1 (i.e., MC items and OE task were all weighted 1), was 0.56. Since the raw score is a sufficient statistic for estimates of ability, all other students with a Composite 1 score of 11 received the same 0.56

[1] One such program is the *Advanced Placement Program* of the College Board. For example, their 1996 Biology Examination consisted of four OE questions, each of which is scored on a 0 to 10 rubric, and 120 dichotomously scored MC items. The composite score weighting for this examination was .75(MC) + 1.50(OE), such that the MC items contributed 60 percent and the OE questions contribute 40 percent to the maximum possible composite score of 150, with each of the OE questions contributing equally (See Table 1, Biology, Educational Testing Service, 1997). Once section scores are converted to composite scores, the Chief Reader sets grade boundaries to convert the

ability score. When the weights are changed, the composite scores become 12.35 and 13.78 for Composites 2 and 3 respectively, and the student's ability estimates jump to 1.33 and 3.45.

composite score to AP grades, i.e., 5, very well qualified,; 4, well qualified; 3, qualified; 2, possibly qualified; and 1, no recommendation.

When rater effects are considered, this student's ability estimates become 0.52, 1.30 and 4.17 for Composites 4, 5, and 6. For other students with identical Composite 1 scores rated by the same rater (such as student #25568), the estimates for the Composite 4, 5, and 6 scores become 0.52, 0.79 and 1.30 respectively. This discrepancy occurs because student #13889 scored a 4 on the OE response when all others in the group scored a 3. As would be expected, the student ability score (theta) remains the same for Composite 1 scores since the part-weights are 1 for the multiple-choice and the open-ended task. However, when the OE task is weighted more than 1 and student #13889 scores higher on the task than the other students, then his/her composite score increases, reflecting in higher ability estimate.

The same is true for students who scored lower on the OE task but had identical overall scores. These students had lower ability estimates than those students with identical Composite 1 scores but who had a higher score on their OE task This can be seen for Students #13896, #15338, and #13594, with identical Composite 1 scores of 9 in Table 8. Student #13896 scored a 1 on the OE task, while the students #15338 and #13594 scored 2 and 3 respectively.

As Table 8 portrays, rater #1117 is considered to be lenient overall. With the exception of student #13889, ability scores for the given set of students are all adjusted downward when the rater's effects are included in the estimation. Student # 13889, however, has his/her ability estimate increased when the OE task is weighted most heavily, indicating that the rater's severity in giving 4 plus scores is compounded by weighting the task heavily.

<u>Insert Table 8, here.</u>

In the one-parameter Rasch model, weighted composite scores, like raw scores, are sufficient statistic in estimating student abilities. Student ability estimates, therefore, will be adjusted upwards or downwards depending on students' total composite scores impacted by the

weights assigned to the open-ended responses. When rater effects are included in the measurement model that has weighted composite scores, rater severity is confounded by the weighted composite scores. As expected, the standard errors for rater severity measurement decrease when weights are increased, however, rater severity estimates may increase or reduce with differently weighted composites (see Table 5).

In conclusion, when weighted scores are used differentially, student ability measures are not only a function of the rater that scores them, but also the items/tasks the students answer correctly. It is imperative that weights be assigned on substantive grounds with an understanding of the consequences of assigning indiscriminate weights.

This research lays the path for replication by a study of simulated data. Additional weighted composite scores could be included in the simulation to account for individual MC item weights that are greater than 1. Further diversity could be acquired in the assignment of composite scores by differentially weighting the MC and OE sections and by using both the one and two parameter IRT models. It would also be interesting to see the impact on student ability estimates under crossed rater conditions and when raters are not homogeneous with respect to the number of students they score under nested conditions. Finally, Hombo's. Thayer's, and Donoghue's (2000) suggestion of using a spiral rater design could be incorporated in the study to analyze the effects with weighted composite scores on student ability estimates.

References

Educational Testing Service. (1997, April). College Board 1996 Advanced Placement Test Analyses, Forms 3SBP. (Educational Testing Service Report No. SR-97-23, Unpublished statistical report). Author: Princeton, NJ.

Engelhard, G. (1996). Evaluating rater accuracy in performance assessments. *Journal of Educational Measurement, 31(1)*, 56-70.

Engelhard, G. (1994). Examining rater errors in the assessment of written-composition with the many-faceted Rasch model. *Journal of Educational Measurement, 31(2),* 93-112.

Hombo, C. M., Thayer, D. T., & Donoghue, J. R. (2000). A simulation study of the effect of crossed and nested rater designs on ability estimation. Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans, LA.

Linacre, J. M. (1989). *A user's guide to FACETS, Rasch measurement program, and Facform, data formatting computer program.* Chicago, IL: MESA Press

Linacre, J. M. (1993). *Many-facet Rasch measurement.* Chicago, IL: MESA Press.

Lunz, M. E., Wright, B. D., & Linacre, J. M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3(4),* 331-345.

Masters, G. N. & Wright, B. D. (1981). *A model for partial credit scoring* (Research Memorandum No. 31). Mesa Psychometric Laboratory, Department of Education, University of Chicago.

Table 1. Summary of Weights and Composites.

Part Score Composition	Maximum Part Score	Part Weight Used in *FACETS*	Maximum Weighted Part Score	Part Percent in Composite
Composites 1, 4				
Multiple-Choice Items	14	1.00000	14	73.7
Open-Ended Score	5	1.00000	5	26.3
Composites 2, 5				
Multiple-Choice Items	14	0.67857	9.5	50.0
Open-Ended Score	5	1.90000	9.5	50.0
Composites 3, 6				
Multiple-Choice Items	14	0.33929	4.75	25.0
Open-Ended Score	5	2.85000	14.25	75.0

Note: Composites 1, 2, and3 do not include Rater effects; Composites 4, 5, and 6 include Rater Effects.

Table 2. Descriptive Statistics for Raw Score Parts and Unadjusted Raw Score Composites.

Variable	N	Mean	SD
Multiple-Choice Total	3,200	10.5	2.6
Open-Ended Score	3,200	2.7	0.9
Composite 1 (75/25)	3,200	13.2	3.0
Composite 2 (50/50)	3,200	12.3	2.8
Composite 3 (25/75)	3,200	11.4	2.9

Table 3. Correlations for Raw Score Parts and Unadjusted Raw Score Composites.

Score/Composite	Multiple-Choice Total	Open-Ended Score	Composite 1 (75/25)	Composite 2 (50/50)	Composite 3 (25/75)
Multiple-Choice Total	1.00	.39	.96	.85	.63
Open-Ended Score		1.00	.62	.82	.96
Composite 1 (75/25)			1.00	.96	.81
Composite 2 (50/50)				1.00	.95
Composite 3 (25/75)					1.00

Table 4.Mean and Standard Deviations for Raw Score Parts and Unadjusted Raw Score Composites (by Rater).

	Rater	Multiple-Choice Total	Open-Ended Score	Composite 1 (75/25)	Composite 2 (50/50)	Composite 3 (25/75)
1	1064	10.6 (2.5)	3.0 (0.9)	13.6 (3.0)	12.9 (2.8)	12.1 (2.9)
2	1092	10.4 (2.6)	3.0 (0.8)	13.4 (3.0)	127 (2.8)	12.0 (2.8)
3	1117	10.0 (2.6)	3.0 (0.8)	12.7 (3.1)	11.8 (2.7)	10.9 (2.6)
4	1181	10.1 (2.7)	2.2 (1.0)	12.4 (3.1)	11.1 (2.9)	9.8 (3.1)
5	1402	10.4 (2.5)	2.8 (0.9)	13.2 (3.0)	12.4 (2.98)	11.5 (3.0)
6	1403	10.3 (2.6)	2.6 (0.7)	13.0 (2.9)	12.0 (2.5)	11.0 (2.4)
7	1407	10.9 (2.4)	2.5 (0.8)	13.4 (2.8)	12.1 (2.6)	10.8 (2.7)
8	1408	11.0 (2.5)	2.6 (0.8)	13.6 (3.0)	12.4 (2.7)	12.2 (2.7)
9	402	10.6 (2.8)	2.9 (0.8)	13.5 (3.2)	12.7 (2.9)	11.8 (2.9)
10	468	10.6 (2.3)	3.0 (0.9)	13.6 (2.8)	12.9 (2.8)	12.1 (3.1)
11	520	10.2 (2.6)	2.6 (0.9)	12.8 (3.0)	11.9 (2.8)	10.9 (2.9)
12	591	10.0 (2.7)	2.7 (0.7)	12.6 (3.0)	11.8 (2.7)	11.0 (2.5)
13	671	11.1 (2.4)	2.9 (0.9)	14.0 (2.7)	13.1 (2.6)	12.1 (2.7)
14	767	9.8 (3.1)	2.4 (0.9)	12.3 (3.6)	11.3 (3.3)	10.3 (3.3)
15	817	10.8 (2.4)	3.2 (1.0)	14.0 (2.9)	13.4 (2.9)	12.8 (3.1)
16	943	11.0 (2.5)	2.8 (0.8)	13.8 (2.9)	12.8 (2.6)	11.7 (2.7)

Note: N = 200 students for each rater.

Table 5. Rater Measures for Composite Scores 4, 5, and 6.

	Rater ID	Composite 4 (75/25) Measure	SE	Composite 5 (50/50) Measure	SE	Composite 6 (25/75) Measure	SE
1	1064	0.70	0.11	0.84	0.08	0.78	0.08
2	1092	0.53	0.11	0.68	0.08	0.65	0.08
3	1117	0.46	0.12	0.60	0.09	0.94	0.08
4	1181	0.02	0.09	-0.11	0.07	-0.19	0.07
5	1402	-0.99	0.10	-1.16	0.08	-1.56	0.07
6	1403	-0.68	0.13	-0.81	0.09	-1.01	0.09
7	1407	1.36	0.12	1.71	0.09	1.98	0.08
8	1408	-0.16	0.13	-0.21	0.10	-0.15	0.09
9	402	-1.03	0.12	-1.20	0.09	-1.58	0.08
10	468	0.01	0.10	-0.05	0.07	-0.01	0.07
11	520	0.28	0.11	0.31	0.08	0.69	0.08
12	591	-0.79	0.13	-0.87	0.10	-1.03	0.09
13	671	0.13	0.11	0.10	0.08	0.27	0.08
14	767	-0.36	0.11	-0.47	0.08	-0.54	0.08
15	817	-0.08	0.10	-0.16	0.07	-0.28	0.07
16	943	0.60	0.11	0.80	0.09	1.05	0.08

Table 6. Range of Mean Square INFIT and OUTFIT Statistics for Rater Measures.

Composite Score	Mean Square INFIT	Mean Square OUTFIT
Composite 4 (75/25)	0.9 – 1.2	0.9 – 1.2
Composite 5 (50/50)	0.5 – 0.7	0.5 – 0.7
Composite 6 (25/75)	0.2 – 0.4	0.2 – 0.3

Table7. Classification of Students with Respect to Quartile Cutpoints.

	Percent of Students Above Cut			Percent of Students Changing Classification	
	Unadjusted	Adjusted	Consistent	Up	Down
Composite 1 vs. 4					
Q1	75	69	94	0	6
Median	54	47	94	0	6
Q3	40	33	93	0	7
Composite 2 vs. 5					
Q1	73	69	95	1	4
Median	47	46	95	2	3
Q3	37	34	94	2	4
Composite 3 vs. 6					
Q1	64	66	95	4	1
Median	56	54	94	2	5
Q3	46	43	91	3	6

Table 8. Results for Two Sets of Students Scored by Rater #1117

		Composites 1 and 4 (75/25 Weights)			Composites 2 and 5 (50/50 Weights)			Composites 3 and 6 (25/75 Weights)		
Student ID	OE Score	Raw Score	Unadj. Theta	Adj. Theta	Raw Score	Unadj. Theta	Adj. Theta	Raw Score	Unadj. Theta	Adj. Theta
13889	4	11	0.56	0.52	12.35	1.33	1.30	13.78	3.45	4.17
25568	3	11	0.56	0.52	11.13	0.86	0.79	11.26	1.44	1.30
13896	1	9	-0.01	-0.05	7.33	-0.45	-0.52	5.56	-1.36	-1.49
15338	2	9	-0.01	-0.05	8.55	-0.03	-0.11	8.08	-0.12	-0.29
13594	3	9	-0.01	-0.05	9.77	0.38	0.30	10.59	1.07	0.90

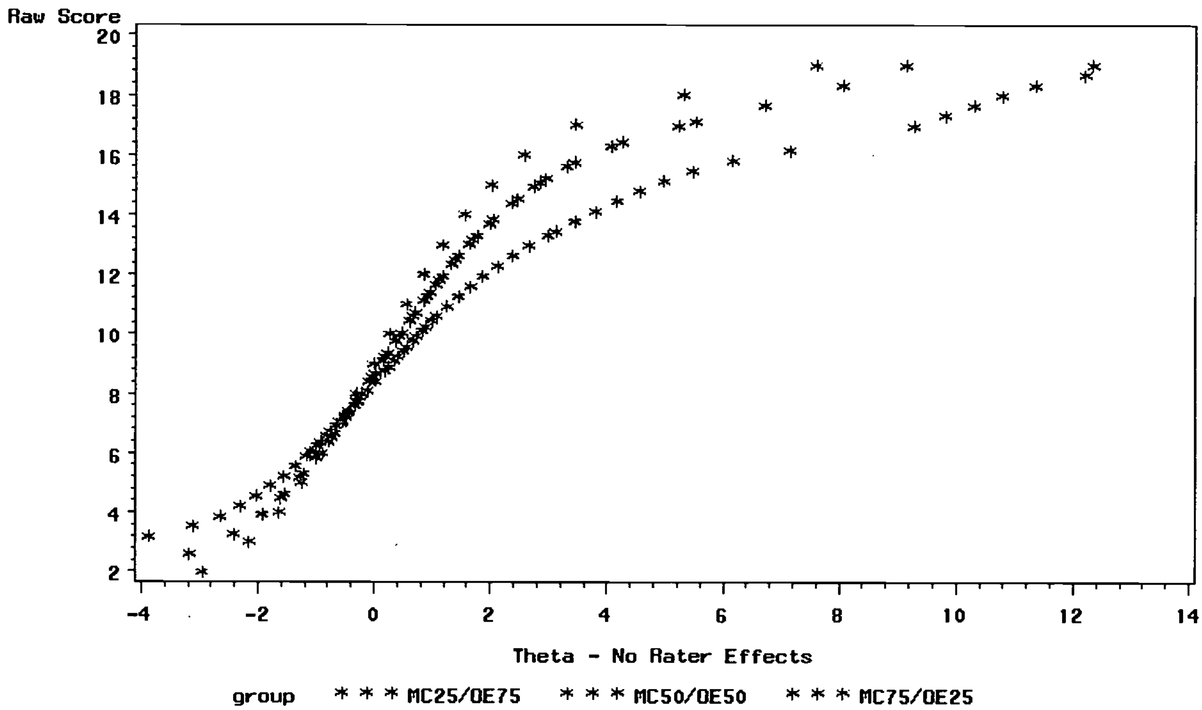

Figure 1. Plots of Raw Score vs. Theta for Composite Scores 1, 2, and 3 (No Rater Effects).

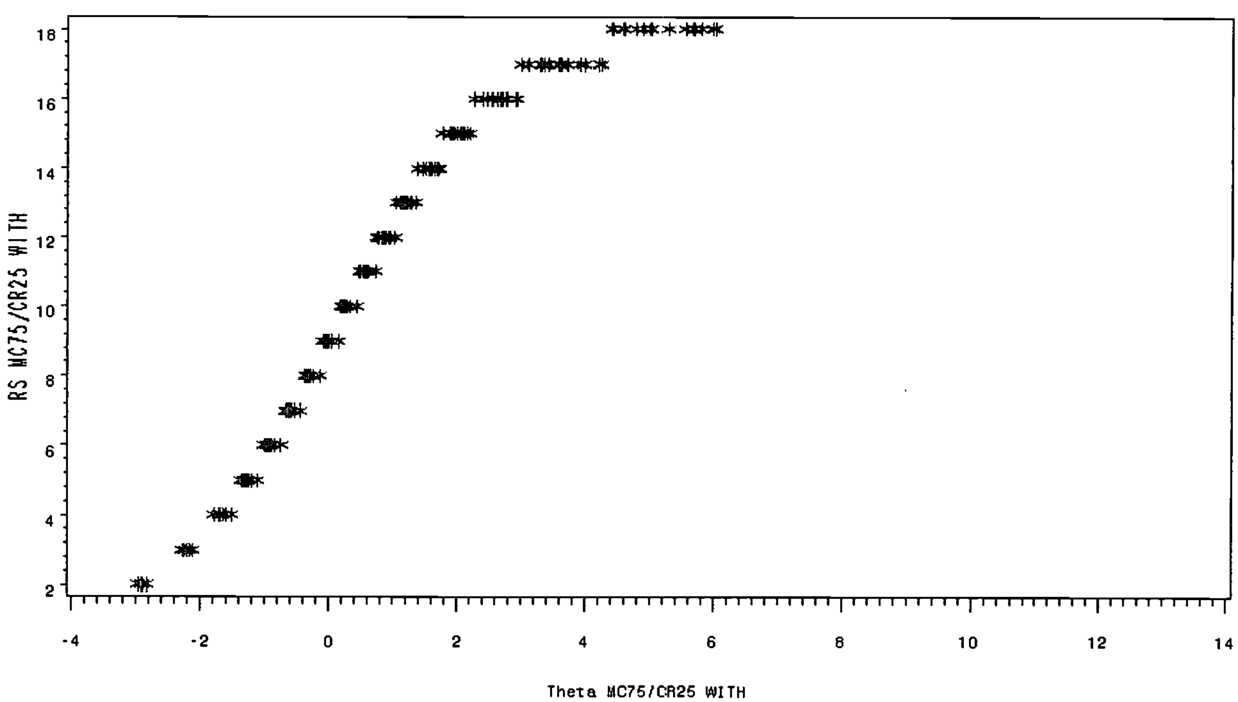

Figure 2. Plot of Raw Score vs. Theta for Composite Score 4 (With Rater Effects).

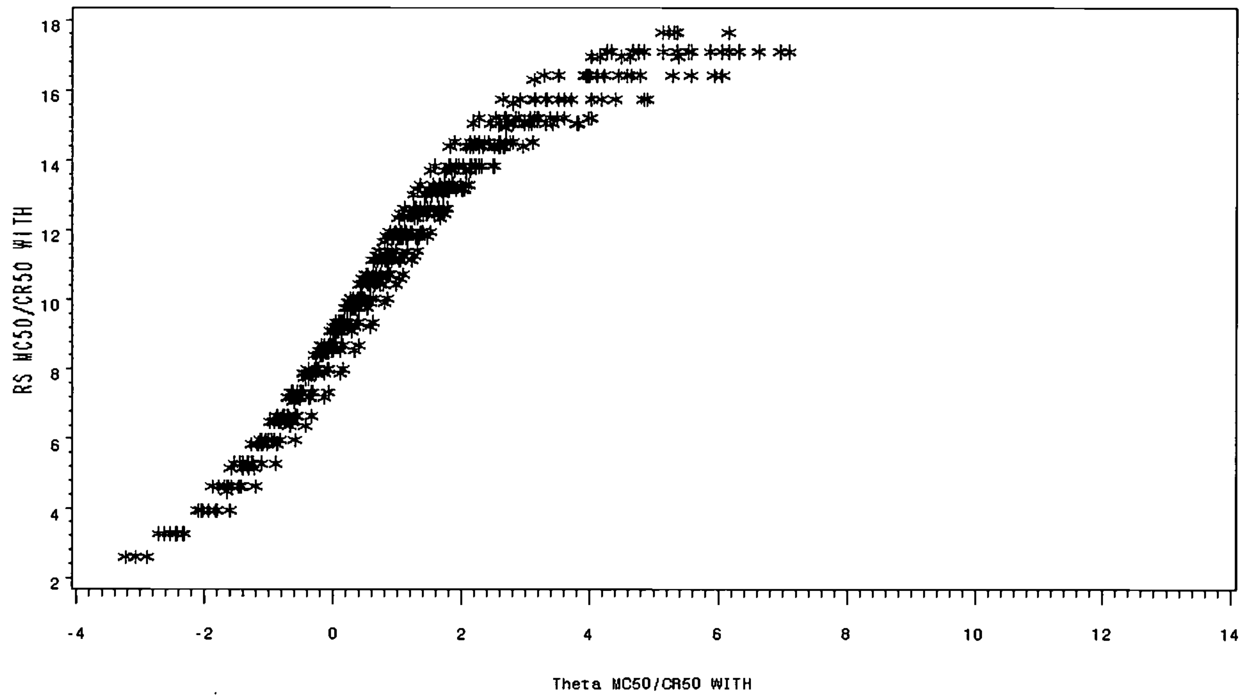

Figure 3. Plot of Raw Score vs. Theta for Composite Score 5 (With Rater Effects).

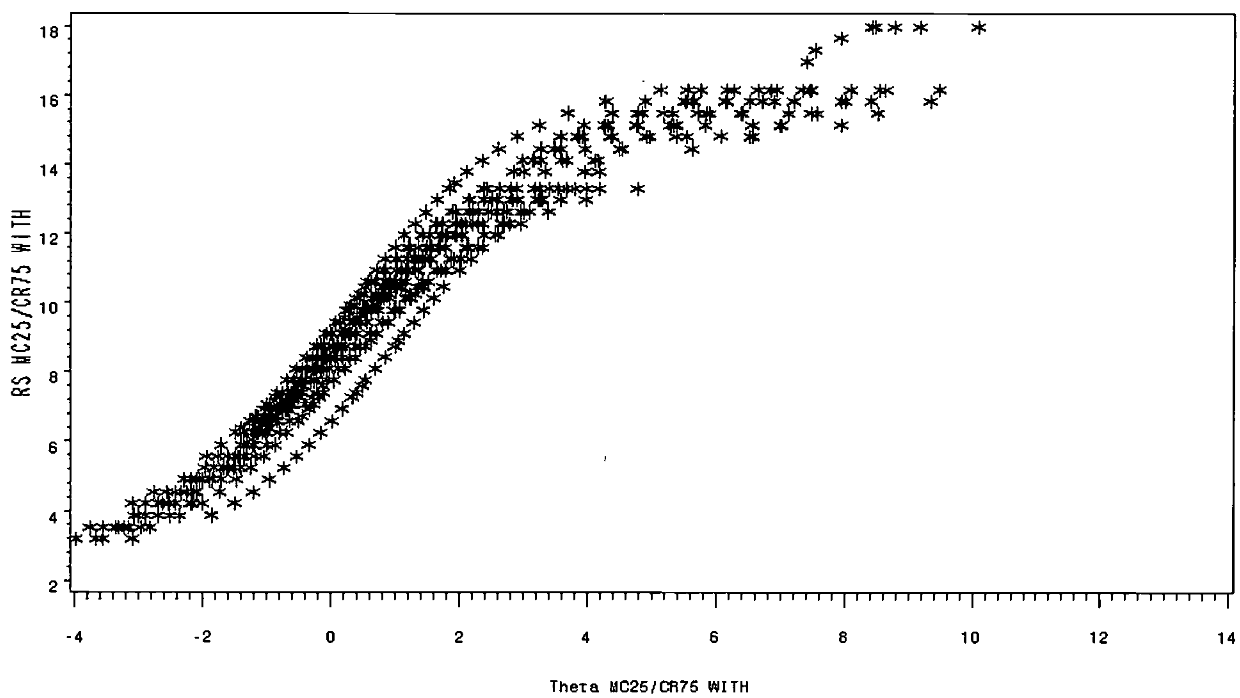

Figure 4. Plot of Raw Score vs. Theta for Composite Score 6 (With Rater Effects).

Sign here, → please

Signature: H. Taherbhai	Printed Name/Position/Title: HUSEIN M. TAHERBHAI	
Organization/Address: LRDC, UNIVERSITY of PITTSBURGH, 3939 O'HARA ST., RM. 802, PITTSBURGH, PA 15260	Telephone: 412-624-7569	FAX:
	E-Mail Address: taherbha@pitt.edu	Date: 4/17/01

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

THE UNIVERSITY OF MARYLAND
ERIC CLEARINGHOUSE ON ASSESSMENT AND EVALUATION
1129 SHRIVER LAB, CAMPUS DRIVE
COLLEGE PARK, MD 20742-5701
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com